The Maximum Entropy Model with Continuous Features

Dong Yu, Li Deng, Alex Acero
Microsoft Research
One Microsoft Way
Redmond, WA 98052
{dongyu, deng, alexac}@microsoft.com

Abstract

We present the maximum entropy (MaxEnt) model with continuous features. We show that for the continuous features the weights should be continuous functions instead of single values. We propose a spline interpolation based solution to the optimization problem that contains continuous weights and illustrate that the optimization problem can be converted into a standard log-linear one without continuous weights at a higher-dimensional space.

1 Introduction

The maximum entropy (MaxEnt) model with moment constraints on binary features has been shown to be effective (e.g., Berger et al. 1996; Yu et al. 2005; Ma et al. 2007). However, it is not as successful when continuous features are used. In the past, people have found that quantization techniques (e.g. bucketing or binning) help sometimes.

In this paper, we present the MaxEnt model with continuous features. We point out that the weights for continuous features in the MaxEnt model should not be single values but continuous functions. We further provide a solution to the optimization problem that contains continuous weights using spline interpolations (Yu et al. 2008) and convert the optimization problem into a standard log-linear optimization problem with only single-value weights at a higher-dimensional space where each original continuous feature is mapped into several features. The existing training and testing algorithms for the MaxEnt model can thus be directly applied to this higher-dimensional space.

The rest of the paper is organized as follows. In Section 2, we examine the MaxEnt model with moment constraints. In Section 3, we show that continuous weighting functions should be used for continuous features and propose a solution to the optimization problem that contains continuous weights. We conclude the paper in Section 4.

2 MaxEnt Model with Moment Constraints

Let us consider a random process that produces an output value y from a finite set Y for some input value x. We assume that a training set $(x_1, y_1), (x_2, y_2), \cdots, (x_N, y_N)$ with N samples is given. Our goal is to construct a stochastic model that can accurately represent the random process that generated the training set. We denote $p(y|x)$ as the probability of outputting y by the model when x is given. The MaxEnt principle dictates that for all the models that confine to the constraints C we should select the model that maximizes the entropy. If only the constraints on the first order moment

$$E_p[f_i] = E_{\tilde{p}}[f_i], i = 1, \cdots, M \tag{1}$$

are used, where

$$E_p[f_i] = \sum_{x,y} \tilde{p}(x) p(y|x) f_i(x, y) \tag{2}$$

and

$$E_{\tilde{p}}[f_i] = \sum_{x,y} \tilde{p}(x,y) f_i(x,y) = \sum_{x,y} \tilde{p}(x) \tilde{p}(y|x) f_i(x,y), \tag{3}$$

The solution to the MaxEnt model is in the log-linear form of (Berger et al. 1996)

$$p(y|x) = \frac{1}{Z_\lambda(x)} \exp\left(\sum_i \lambda_i f_i(x,y) \right), \tag{4}$$

where

$$Z_\lambda(x) = \sum_y \exp\left(\sum_i \lambda_i f_i(x,y) \right) \tag{5}$$

is a normalization constant to make sure $\sum_y p(y|x) = 1$, and λ is chosen to maximize

$$\Psi(\lambda) = -\sum_x \tilde{p}(x) \log Z_\lambda(x) + \sum_i \lambda_i E[\tilde{p}(f_i)]. \tag{6}$$

Typical algorithms used to solve the above convex optimization problem include generalized iterative scaling (GIS) (Darroch & Ratcliff 1972) and gradient ascent and conjugate gradient (e.g. L-BFGS) (Nocedal 1980) based algorithms.

With the binary features where $f_i(x,y) \in \{0,1\}$, the moment constraint (1) is a strong constraint since $E_p[f] = p(f = 1)$. However, the moment constraint is rather weak for continuous features.

3 MaxEnt Model with Continuous Features

To get a better statistical model, quantization techniques such as bucketing (or binning) have been proposed to convert the continuous features to the binary features, with which a continuous feature f_i in the range of $[l, h]$ can be converted into K binary features

$$f_{ik}(x,y) = \begin{cases} \dfrac{h_k + l_k}{2} & if\ y\ matches\ and\ x \in [l_k, h_k] \\ 0 & otherwise \end{cases} \tag{7}$$

where $k \in \{1,2,\cdots,K\}$, and $l_k = h_{k-1} = (k-1)(h-l)/K + l$. Using bucketing we essentially approximate the constraints on the distribution of the continuous features with the moment constraints on each segment.

Now assume we have infinite number of training samples and we may increase the number of buckets to any large number we want. Under this condition, we have

$$\lim_{k\to\infty} \sum_k \lambda_{ik} f_{ik}(x,y) = \lambda_i\big(f_i(x,y)\big) f_i(x,y) \tag{8}$$

by noting that only one $f_{ik}(x,y)$ is none-zero for each (x,y) pair, where $\lambda_i\big(f_i(x,y)\big)$ is a continuous weighting function over the feature values. (8) suggests that when continuous features are used the solution to the MaxEnt model is

$$p(y|x) = \frac{1}{Z_\lambda(x)} \exp\left(\sum_{i \in \{continuous\ \}} \lambda_i\big(f_i(x,y)\big) f_i(x,y) + \sum_{j \in \{binary\ \}} \lambda_j f_j(x,y) \right) \tag{9}$$

which contains continuous weights. Let us further approximate each continuous weight using a cubic-spline with natural boundary conditions. As we have shown in Yu et al. (2008) that given K evenly distributed knots $\{(f_{ik}, \lambda_{ik}) | k = 1,\cdots,K\}$ where $h = f_{ik+1} - f_{ik} = f_{ij+1} - f_{ij} > 0, \forall j,k \in \{1,\cdots,K-1\}$, $\lambda_i(f_i)$ can be approximated as

$$\lambda_i(f_i) \cong \boldsymbol{a}^{\boldsymbol{T}}(f_i)\boldsymbol{\lambda}_i \tag{10}$$

where $\boldsymbol{\lambda}_i = [\lambda_{i1}, \cdots, \lambda_{iK}]^T$ and $\boldsymbol{a}^{\boldsymbol{T}}(f_i)$ is a vector (Yu et al. 2008). Note that with (10) we have

$$\lambda_i(f_i)f_i \cong \boldsymbol{a}^{\boldsymbol{T}}(f_i)\boldsymbol{\lambda}_i f_i = [\boldsymbol{a}^{\boldsymbol{T}}(f_i)f_i]\boldsymbol{\lambda}_i = \sum_k \lambda_{ik}[a_k(f_i)f_i], \tag{11}$$

where $a_k(f_i)$ is the k-th element of $\boldsymbol{a}^{\boldsymbol{T}}(f_i)$. (11) indicates that the optimization problem (9) can be converted into

$$p(y|x) = \frac{1}{Z_\lambda(x)} \exp\left(\sum_{i \in \{continuous\ \},k} \lambda_{ik} f_{ik}(x,y) + \sum_{j \in \{binary\ \}} \lambda_j f_j(x,y) \right), \tag{12}$$

where

$$f_{ik}(x,y) = a_k\big(f_i(x,y)\big) f_i(x,y). \tag{13}$$

(12) is in the standard log-linear form at a higher-dimensional space and can be solved with existing algorithms that supports negative values.

To validate our theory, we have compared it with the single-constraint MaxEnt model where each continuous feature is constrained only on its mean, and the bucketing approach where each continuous feature is quantized into K segments, on two classification tasks from the UCI data repository (Asuncion & Newman 2007). In all the experiments, our proposed approach consistently outperforms the single-constraint MaxEnt model and the bucketing approach with significant margins.

4 Summary and Discussion

In this paper, we presented the MaxEnt model with continuous features. We showed that for continuous features, the weights should be continuous functions instead of single values. We provided a solution to the optimization problem that contains continuous weights. The beauty of our solution is that we can spread and expand each original feature into several features at a higher-dimensional space through a non-linear mapping. With this feature conversion, the optimization problem with continuous weights is transformed into a standard log-linear feature combination problem and the existing MaxEnt algorithms can thus be directly used.

References

A. Asuncion, & D. J. Newman. (2007). UCI Machine Learning Repository [http://www.ics.uci.edu/~mlearn/MLRepository.html]. Irvine, CA: University of California, School of Information and Computer Science.

A. L. Berger, S. A. Della Pietra, & V. J. Della Pietra. (1996) A Maximum Entropy approach to Natural Language Processing. *Computational Linguistics*, vol. 22, pp. 39-71.

J. Darroch & D. Ratcliff. (1972) Generalized iterative scaling for log-linear models. *Ann. Math. Statistics*, 43:1470–1480

C. Ma, P. Nguyen, and M. Mahajan. (2007) Finding Speaker Identities with a Conditional Maximum Entropy Model. In proc. of ICASSP 2007, pp. IV-261-IV-264.

M. Mahajan, A. Gunawardana, & A. Acero. (2006) Training Algorithms for Hidden Conditional Random Fields, in proc. of ICASSP 2006, pp. I-273-I-276.

J. Nocedal. Updating Quasi-Newton Matrices with Limited Storage (1980), *Mathematics of Computation*, vol. 35, pp. 773-782.

D. Yu, M. Mahajan, P.Mau, & A. Acero. (2005) Maximum Entropy Based Generic Filter for Language Model Adaptation, in proc. of ICASSP 2005, pp. I-597-I-600.

D. Yu, L. Deng, Y. Gong, & A. Acero. (2008) Discriminative Training of Variable-Parameter HMMs for Noise Robust Speech Recognition, in proc. of Interspeech 2008, pp. I-285-I-288.